REMARKS

Claims 1-31, 35-44, 46-48, 52, and 53 are pending. Claims 1, 9, 17, 18, 31 35, 48, 52, and 53 have been amended. Claims 7, 8, 15, 16, 28-30, 46, and 47 have been canceled. No new matter has been entered. Claims 1-6, 9-14, 17-27, 31,

5   35-44, 48, 52, and 53 remain.

Claims 1-23, 28, 29, 35-40, 46, 52, and 53 stand rejected under 35 U.S.C. § 103(a) as being obvious in light of International Application Publication No. WO 03/060766, to Lindh et al. ("Lindh"), in view of U.S. Patent No. 6,560,597, issued to Dhillon et al. ("Dhillon"). Applicant traverses.

10   One of ordinary skill in the art at the time of applicant's invention would not have been motivated or have had a reason to combine Lindh and Dhillon. Lindh teaches preprocessing a corpus of documents stored in a database, including performing word splitting, identifying proper names, removing stop words, applying a word stemming algorithm, and performing word weightings (p.

15   19, lines 2-5). Following preprocessing, each unique term is assigned a weight according to that term's information content, which is determined using a Term Frequency times Inverse Document Frequency (TFIDF) (p. 17, lines 21-23). Matrices are generated to describe relationships within the document corpus using the unique terms (p. 18, lines 23-25). Lindh further teaches enhancing

20   relationship quality by filtering the document corpus to generate a term-term matrix (p. 27, lines 18-25). A reduction in the number of similar documents in the corpus precludes large quantities of similar documents from biasing the relationship measures, which is characterized as a flaw that can be reduced using document clustering, such as k-means clustering (p. 27, line 25-p. 28, line 5). A

25   representative document vector is generated for each cluster found by a clustering algorithm, such as by calculating a cluster centroid as the mean of all document vectors in the cluster (p. 28, lines 8-23). The representative document vector is added to the cluster and all other documents that belong to the cluster are removed from the initial document corpus (p. 28, lines 8-23).

30   In contrast, Dhillon teaches defining document concept decomposition

vectors that represent a document vector space (Abstract). Documents, which are received from a text repository (Col. 3, lines 53-55), are parsed to remove redundant words for determining word frequency counts (Col. 4, lines 13-19). A disjoint clustering of the documents is performed by dividing the document vector

5   space into partitions (Col. 5, lines 9-23). An initial set of concept vectors is determined as a mean vector of all document vectors for each partition (Col. 5, lines 56-60). A new partitioning of the document vector space is formed after the determination of the initial concept vectors (Col. 6, lines 288-30). New concept vectors corresponding to the new partitions are then formed (Col. 6, lines 37-45).

10   The partitioning continues until stop criteria are met, such as a magnitude of change or a predetermined number of iterations (Col. 6, lines 46-58). Once stopped, the document vectors are projected into a subspace (Col. 6, lines 59-65).

Independent Claims 1, 9, 17, 18, 35, 52, and 53 have been amended. Claim 1 has been amended to include the limitations of now-canceled Claims 7, 8,

15   and 29. Amended Claim 1 recites a seed document identification submodule identifying a set of seed documents by applying the similarity as a best fit to each such candidate seed document. Claim 1 further recites a clustering submodule grouping each such non-seed document into a cluster with the best fit, subject to a minimum fit. Claim 9 has been amended to include the limitations of now-

20   canceled claims 15, 16, and 46. Amended Claim 9 recites identifying a set of seed documents by applying the similarity as a best fit to each such candidate seed document. Claim 9 further recites grouping each such non-seed document into a cluster with the best fit, subject to a minimum fit. Claim 17 has been amended to include the limitations consistent with amended Claim 1, which now recites code

25   for identifying a set of seed documents by applying the similarity as a best fit to each such candidate seed document. Claim 17 further recites code for grouping each such non-seed document into a cluster with the best fit, subject to a minimum fit.

Similarly, Claims 18, 35, 52, and 53 have been amended. Claim 18 has been amended to include the limitations of now-canceled Claims 7, 8, and 29. Claim 18 recites a cluster seed submodule identifying seed documents by applying the similarity as a best fit to each such candidate seed document. Claim

5 18 further recites a clustering submodule assigning each non-seed document to the cluster with the best fit, subject to a minimum fit. Claim 35 has been amended to include the limitations of now-canceled Claims 15, 16 and 46. Claim 35 recites identifying seed documents-by applying the similarity as a best fit to each such candidate seed document. Claim 35 further recites assigning each non-seed

10 document to the cluster with the best fit, subject to a minimum fit. Claims 52 and 53 have been amended to include the limitations consistent with amended Claim 18. Claim 52 recites code for identifying seed documents by applying the similarity as a best fit to each such candidate seed document. Claim 52 further recites code for assigning each non-seed document to the cluster with the best fit,

15 subject to a minimum fit. Claim 53 recites means for identifying seed documents by applying the similarity as a best fit to each such candidate seed document. Claim 53 further recites means for assigning each non-seed document to the cluster with the best fit, subject to a minimum fit.

The claim amendments should not necessitate a new ground of rejection

20 based on prior art not of record, as each of the limitations in the claim amendments were already considered and examined in the previous Office action. *See* MPEP 706.07(a) ("A second or any subsequent action on the merits in any application or patent involved in reexamination proceedings should not be made final if it includes a rejection, on prior art not of record, of any claim amended to

25 include limitations which should reasonably have been expected to be claimed" (emphasis added)).

The combination of Lindh and Dhillon fail to render Claims 1, 9, 17, 18, 35, 52, and 53 obvious. Per Lindh, the combined references disclose document clustering via a clustering algorithm to reduce bias of relationship values due to

30 large quantities of similar documents (Lindh, p. 27, line 25-p. 27, line 3; p. 28,

lines 9-11). A representative document vector is generated for each cluster by calculating a centroid, which is the mean of all document vectors in that cluster (Lindh, p. 28, lines 11-14). Once determined, the representative document vector is added to the corresponding cluster (Lindh, p. 28, lines 14-16). The

5   representative document vector represents the mean value of <u>all document vectors</u> in a particular <u>cluster</u>, instead of a value for a single document. Thus, one representative document vector is assigned to each cluster, which has already been identified via the clustering algorithm. Therefore, the combination teaches assigning a mean document vector to each <u>cluster</u>, rather than identifying <u>seed</u>

10   <u>documents</u> by applying the similarity as a best fit to each candidate seed document.

Further, the Lindh-Dhillon combination fails to teach assigning non-seed documents into a cluster with a best fit, subject to a minimum fit. Documents can be clustered using a clustering algorithm, such as k-means clustering (Lindh, p.

15   28, lines 9-11). A set of clusters containing similar documents will be produced (Lindh, p. 28, lines 6-7). Thus, each document will be clustered with similar documents based on a particular algorithm without applying further requirements, such as a minimum fit criterion. Therefore, the combination teaches assigning documents to clusters using a clustering algorithm, rather than grouping a non-

20   seed document into a cluster with a best fit, <u>subject to a minimum fit</u>.

Moreover, per Dhillon, the combination discloses performing a disjoint clustering of a preprocessed document space (Dhillon, Col. 5, lines 12-15). During the clustering, the document vector space is divided into a uniform distribution of equal-sized partitions along a document dimension index, or by

25   arbitrarily selecting a number of partitions and randomly assigning documents to partitions (Dhillon, Col. 5, lines 16-20). For each partition, an initial concept vector representing a mean vector of all documents in that partition is determined (Dhillon, Col. 5, lines 56-60). Next, a new partitioning of the document vector space is formed by finding a closest concept vector for each document vector

30   (Dhillon, Col. 6, lines 28-37). Each document is placed in a partition based on a

relationship between the document vector and the concept vector. The relationship requires the concept vector to be the <u>closest</u> to the document vector. As long as the concept vector is the closest, the document will be placed in the corresponding partition without having to satisfy further requirements. Thus, the

5   Lindh-Dhillon combination teaches placing documents based on a <u>closest</u> relationship to a concept vector, rather than grouping a non-seed document into a cluster with a best fit, <u>subject to a minimum fit</u>.

Further, the combination teaches iterative partitioning of documents, which results in a concept vector subspace of document vectors (Dhillon, Col. 3,

10   line 64-Col. 4, line 2). During partioning iterations, a closest concept vector is found for each document vector to create new partitions (Dhillon, Col. 6, lines 31-37). New concept vectors are determined for the new partitions after each partitioning iteration (Dhillon, Col. 6, lines 37-40). Partition quality is determined by an objective function that measures the combined coherence of all

15   clusters in a partition (Dhillon, Col. 7, lines 37-41). Quality values of the new and previous partitions are compared to determine a change in quality, which is based on a change between a previous grouping of documents and a new grouping of documents using mean concept vectors (Dhillon, Col. 7, lines 64-67). The change is applied to a stop threshold, which is determined by a *predetermined static*

20   *value*, such as a magnitude of change in a concept vector from one iteration to another or number of iterations (Dhillon, Col. 6, lines 48-51; Col. 7, lines 64-67). If the threshold is met, the partitioning iterations will stop; otherwise, the partitioning iterations continue. Thus, the combination teaches comparing mean concepts of a partition to determine a value for applying a <u>predetermined</u>

25   threshold, rather than <u>dynamically</u> determining a threshold for each cluster as a function of the similarities between the documents grouped into each cluster based on the center of the cluster and the scores assigned to each of the concepts.

Moreover, modifying the teachings of Lindh to consider dynamic data would not be predicable, as Dhillon teaches a <u>static</u> threshold for determining a

30   stopping point for partitioning iterations. A fixed threshold is not adaptable, and

replacing the fixed threshold with a dynamic threshold requires implementing functionality that continually adapts the threshold. Dhillon neither teaches nor suggests allowing the threshold to be dynamically redefined, per Claims 1, 9, 17, 18, 35, 52, and 53.

5   Accordingly, a *prima facie* case of obviousness has not been shown. Claims 2-6 are dependent on Claim 1 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 10-14 are dependent on Claim 9 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 19-23 are dependent on

10   Claim 18 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 36-40 are dependent on Claim 35 and are patentable for the above-stated reasons, and as further distinguished by the limitations therein. Claims 7, 8, 15, 16, 28, 29, and 46 have been canceled. Withdrawal of the rejection is requested.

15   Claims 24-27 and 41-44 stand rejected under 35 U.S.C. § 103(a) as being obvious over Lindh and Dhillon as applied to Claims 18 and 35 above, and further in view of U.S. Patent No. 6,675,159, issued to Lin et al. ("Lin"). Applicant traverses.

Claims 24-27 are dependent upon Claim 18 and are patentable for the

20   reasons stated above, and as further distinguished by the limitations therein. Claims 41-44 are dependent upon Claim 35 and are patentable for the reasons stated above, and as further distinguished by the limitations therein. Withdrawal of the rejection is requested.

Claims 30, 31, 47, and 48 stand rejected under 35 U.S.C. § 103(a) as being

25   obvious over Lindh and Dhillon as applied to Claim 29 above, and further in view of Lin. Applicant traverses.

Claim 31 is dependent upon Claim 18 and is patentable for the reasons stated above, and as further distinguished by the limitations therein. Claim 48 is dependent upon Claim 35 and is patentable for the reasons stated above, and as

30   further distinguished by the limitations therein. Claims 30 and 47 have been

- 23 -

canceled. Withdrawal of the rejection is requested.

Claims 1-6, 9-14, 17-27, 31, 35-44, 48, 52, and 53 are believed to be in condition for allowance. Entry of the foregoing remarks is requested. Reconsideration of the claims and a Notice of Allowance are earnestly solicited.

5   Please contact the undersigned at (206) 381-3900 regarding any questions or concerns associated with the present matter.

Respectfully submitted,

10
Dated: August 31, 2007   By: _____

Krista A. Wittman, Esq.
Reg. No. 59,594

15   Cascadia Intellectual Property
500 Union Street, Ste 1005   Telephone: (206) 381-3900
Seattle, WA 98101   Facsimile: (206) 381-3999

OA 2 Resp
20